**Final Report**: **Olympic Games**

**Team members:**

Ariel

Irina

Portia

**Motivation**

It is the season for Olympic Games. Different athletes from different countries compete for the event's prestigious medals. Besides the hype about the event, host cities also use the games to showcase their food and culture to attract tourists during the games and beyond! Our team was curious to discover what data is available on the subject besides athletes, countries, sports and medals. , and to discover whether host countries have won more medals.

**Project description/outline**

Hypothesis: Host countries have won more medals, in our case study we will focus on the USA.

This project will consider datasets on where the Olympic Games held, athletes, participating countries, costs and medals won.

The project outline is as follows:

1. **Finding Data**

Our team data files were obtained from the following two sources of data: [data.world](<https://data.world/>) and [Kaggle](https://www.kaggle.com

2. **Extracting Files into DataFrames**

\* Extracted CSV files were inspected to identify possible flaws including: different column titles, labelling format, and dates.

\* Files were saved under the "Resources" folder.

\* On Jupyter, dependencies were imported dependencies.

\* A pathway for the csv files was created and the csv files into DataFrames.

3. **Transforming DataFrames**

* Athletes DataFrame

\* Created a Medals DataFrame and split DataFrame to three using groupby function by medal categories: Gold, Silver, and Bronze.

\* The three DataFrames were concatenated into a single DataFrame.

\* Additional columns for medals count were inserted.

\* Medals and athletes tables were merged. Columns were filtered, renamed and formatted. Also columns with empty rows and duplicate insertions were dropped.

\* Designated a unique id

\* Changed the language for several cities to English

* **Costs DataFrames**

\* Filtered and renamed columns.

\* Standardized NaN values

\* Merged the two cost DataFrames based on year and city.

* **Created Unique DataFrames**

\* Created  DataFrames with a list of unique cities and years in order to transfer to the DataBase.

4. **Data Loading**

\* Used an ERD tool to depict relationships between Tables.

\* Initialized engine to connect to postgress (pgAdmin). Created tables, loaded data and joined them to get the relevant columns to our question. This database was chosen because the Olympics is a very structured event, and we believed that a relational database would be the best way to load the data. The data was structured in a way, that there were many common columns that we could use to merge the data together.